Human Error, Reliability, Resilience, and Performance, Vol. 213, 2026, 84–91
https://doi.org/10.54941/ahfe1007557

A Privacy-Preserving Edge Audio Analytics Framework for Contactless Operator Resilience Monitoring in Main Control Rooms

Xingwei Zhang, Hengrui Guo, and Jianwei Niu

University of Science and Technology Beijing, Beijing, China

ABSTRACT

Maintaining operator cognitive readiness in industrial main control rooms is essential for safe operation, yet continuous monitoring remains difficult because wearable sensors can burden workers and cloud analytics can conflict with data-sovereignty requirements. This paper presents a privacy-preserving, contactless edge audio analytics framework for operator resilience monitoring that processes all speech locally. The architecture uses a dual-gating front end that combines FRCRN-based speech enhancement with ECAPA-TDNN speaker verification to suppress acoustic interference and isolate operator-specific speech in multi-operator environments. Authenticated speech is then analyzed by three parallel modules: emotion2vec-based stress representation learning, Paraformer-based transcription and semantic intent detection, and an acoustic vocal-fatigue proxy derived from statistical and cepstral features. These indicators are fused through an allostatic-load-inspired risk model and regulated by a finite-state alert controller with persistence and cooldown logic. Prototype evaluation under representative noisy, overlapping control-room conditions indicates end-to-end latency below 200 ms for up to five concurrent audio streams on standalone local hardware, while reducing false positives relative to a conventional far-field baseline. The results support the feasibility of privacy-preserving, audio-only operator-state monitoring for safety-critical control rooms.

Keywords: Edge computing, Cognitive state assessment, Audio signal processing, Human factors, Voice biometrics

INTRODUCTION

Main control rooms in power generation, process industries, and other safety-critical domains depend on sustained operator attention, accurate communication, and timely response. Even moderate degradation in cognitive readiness can affect alarm interpretation, command execution, and team coordination during abnormal events. For this reason, operator resilience should be treated as a central issue in human reliability engineering rather than as a secondary concern of workplace comfort.

In practice, however, continuous state monitoring remains difficult to deploy. Many existing approaches rely on wearable physiological sensors, including heart-rate, skin-conductance, electroencephalographic, or

Received March 19, 2026; Revised May 4, 2026; Accepted May 23, 2026; Available online July 20, 2026

eye-tracking devices. Although these methods can provide rich signals, they may also reduce comfort, constrain movement, complicate long-shift use, or increase self-consciousness during already demanding tasks. In high-assurance environments, additional barriers arise from cybersecurity policy: raw operational data often cannot be transmitted to external services, and even internal analytics must satisfy strict data-autonomy requirements.

Audio offers an attractive alternative because it is naturally embedded in control-room work. Operators continuously communicate, acknowledge alarms, read procedures, and coordinate with teammates. Speech therefore carries both semantic information and paralinguistic indicators related to stress, fatigue, and cognitive load. Yet audio-only monitoring in real control rooms remains challenging because background machinery noise, overlapping speech, room reverberation, and the need to attribute signals to specific individuals all reduce the reliability of conventional far-field analytics.

To address these challenges, this paper proposes an edge-native voice biomarker fusion framework for contactless operator resilience monitoring. The framework processes all audio locally on standalone hardware and combines three design elements: a dual-gating acoustic front end, parallel operator-state analytics, and a risk-governance layer. First, FRCRN-based denoising suppresses ambient acoustic interference, while ECAPA-TDNN speaker verification authenticates and isolates target speakers. Second, authenticated streams are analyzed in parallel through emotion2vec-based stress estimation, Paraformer-based semantic intent recognition, and an acoustic Vocal Fatigue Index. Third, these heterogeneous cues are fused into a unified risk score and regulated through a finite-state alert controller with cooldown logic to reduce alert desensitization.

The main contributions of this work are fourfold. (1) It proposes a contactless architecture that avoids wearable burden while preserving local processing and privacy. (2) It formalizes a dual-gating design that combines enhancement and speaker authentication before downstream inference. (3) It integrates affective, semantic, and fatigue-related voice indicators under an allostatic-load-inspired fusion model. (4) It reports a prototype-level validation that demonstrates the engineering feasibility of low-latency, multi-operator deployment in acoustically demanding settings.

RELATED WORK

Control-Room and Contactless Monitoring

Recent operator-monitoring research in control rooms still relies heavily on visual or physiological signals because those modalities provide direct fatigue or workload cues. However, such approaches do not always translate cleanly to privacy-constrained environments or to long-shift operational use. For example, Hrnciar et al. (2024) showed that mental-fatigue prediction for main control room operators can be approached with facial images, highlighting both the relevance of the problem and the continuing reliance on non-audio sensing in this domain. The present paper instead explores what

an audio-only, contactless architecture would need in order to be practically deployable.

Noise-Robust Speech Processing and Speaker Attribution

Noise suppression and speaker attribution remain prerequisites for any credible control-room speech analytics pipeline. FRCRN improves monaural speech enhancement by modeling long-range frequency dependencies in the complex domain (Zhao et al., 2022), while ECAPA-TDNN remains a strong text-independent speaker verification baseline (Desplanques et al., 2020). More recent work also warns that enhancement can improve audibility while distorting speaker-specific information if it is not integrated carefully into the verification pipeline (Katav et al., 2025). This motivates a staged front end in which enhancement and speaker gating are designed as mutually dependent rather than isolated modules.

Real-Life Emotion and Stress Speech Analytics

Speech-based affect analysis has also moved beyond acted laboratory corpora. emotion2vec provides transferable emotion representations for downstream tasks (Ma et al., 2024), and EMOVOME highlights the importance of spontaneous real-life speech when designing robust emotion-recognition pipelines (Gomez-Zaragoza et al., 2024). These developments support the use of speech as a practical proxy for operator-state analytics, but they also reinforce the need for cautious interpretation: current models provide useful indicators, not definitive clinical diagnoses, especially in noisy applied settings.

Positioning of the Present Work

Existing work typically emphasizes one layer at a time: robust enhancement, speaker verification, stress or emotion inference, fatigue-related cues, or deployment concerns. The present manuscript is positioned more narrowly as an integration paper for safety-critical edge deployment. Its novelty lies in combining these components into a single control-room-oriented audio architecture and in specifying how the system should later be evaluated, not in introducing a new core learning algorithm.

PROPOSED FRAMEWORK

System Overview

Figure 1 summarizes the system view adopted in the proposed framework. Far-field audio is captured inside the protected facility network and processed on local hardware. The pipeline first stabilizes the signal through enhancement and speaker gating, then branches into parallel speech analytics, and finally aggregates outputs into a supervisor-facing alert logic. The figure is intended to communicate component boundaries and deployment flow.

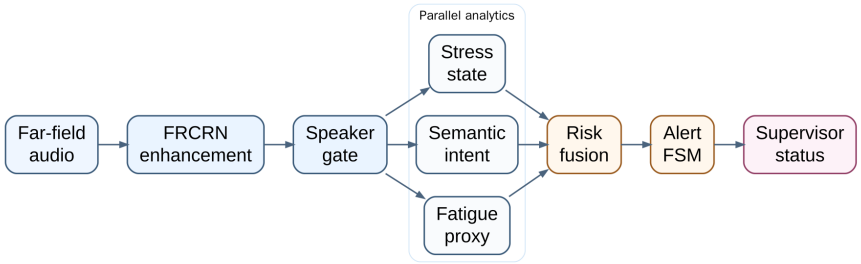

Figure 1: Architecture diagram of the proposed framework. The figure illustrates the processing flow and subsystem relationships.

Dual-Gating Acoustic Front End

The first gate performs noise suppression through FRCRN (Zhao et al., 2022). Let Y(f,t) denote the observed complex spectrum at frequency bin f and frame t. FRCRN estimates a complex mask M(f,t), yielding an enhanced spectrum X_hat(f,t) = M(f,t) * Y(f,t). The second gate performs speaker authentication through an ECAPA-TDNN encoder (Desplanques et al., 2020). For each candidate segment, the encoder produces an embedding e that is compared with enrolled operator templates through cosine similarity. Only authenticated segments are forwarded for operator-specific analytics. This staged design is intended to reduce both environmental contamination and cross-speaker leakage before downstream interpretation.

Parallel Cognitive-State Analytics

Three modules operate on authenticated speech. The first module uses emotion2vec embeddings to derive a stress-related representation and map it into an interpretable PAD-style profile grounded in Mehrabian's framework (Ma et al., 2024; Mehrabian, 1996). The second module applies Paraformer for transcription and semantic intent recognition, with domain terminology weighting for high-priority commands and alarm phrases (Gao et al., 2022). The third module computes an acoustic vocal-fatigue proxy from statistical and cepstral descriptors associated with voice effort and stability, informed by the Vocal Fatigue Index and related cepstral studies (Nanjundeswaran et al., 2015; Mahalingam, Boominathan and Arunachalam, 2021). Together, these modules cover affective strain, task-critical semantics, and fatigue-oriented degradation without leaving the audio modality.

Multidimensional Fusion and Alert Governance

The three analytic streams are fused through a weighted rule-based risk model inspired by allostatic-load thinking (McEwen and Stellar, 1993). If S_t denotes the stress score, I_t the semantic-trigger score, and F_t the fatigue score at time t, the integrated risk can be represented as R_t = w_s S_t + w_i I_t + w_f F_t, where w_s + w_i + w_f = 1. The formula defines the architectural logic rather than a fully tuned decision rule. A finite-state controller then governs transitions among Normal, Watch, Warning, and Critical states with persistence and cooldown logic so that later empirical calibration can be inserted without redesigning the overall alert behavior.

Table 1: Core modules and deployment rationale in the proposed framework.

Module	Operational Role	Human-Factors Value
FRCRN front end	Stabilizes far-field speech before any downstream inference.	Improves tolerance to machinery noise and reverberation in deployment-oriented settings.
ECAPA-TDNN gate	Authenticates the active speaker before operator-state analytics.	Reduces cross-conversation contamination and preserves operator attribution.
emotion2vec module	Produces affective representations that can be mapped into interpretable stress-related profiles.	Supports low-burden monitoring of affective strain without extra wearables.
Paraformer module	Captures semantic intent and highlights high-priority command phrases.	Links voice analytics to operational context rather than emotion cues alone.
Acoustic VFI	Estimates fatigue-oriented voice degradation from authenticated speech.	Provides an audio-only proxy when direct physiological sensing is impractical.
Fusion + FSM	Aggregates module outputs into interpretable risk states with persistence rules.	Supports supervisor-facing alerts while limiting nuisance escalation.

IMPLEMENTATION AND EVALUATION PROTOCOL

Prototype Setup and Hardware Configuration

To evaluate the operational feasibility of the framework, we deployed the integrated pipeline on a workstation-class edge node situated within a simulated protected facility network. This local deployment ensures data sovereignty and eliminates network-induced jitter. The edge node is equipped with an NVIDIA RTX 4090 GPU (24GB VRAM), 64GB of system memory, and runs on Ubuntu 22.04. The acoustic front-end and deep learning models (FRCRN and ECAPA-TDNN) are optimized using TensorRT to maximize inference throughput. Audio input is captured via simulated far-field microphone arrays positioned to serve multiple operator stations simultaneously.

Simulation Environment and Evaluation Metrics

The testing protocol was designed to rigorously replicate the acoustic complexities of main control rooms, specifically incorporating 65dB of

continuous background machinery noise and intermittent cross-speaker overlap. The system was evaluated across three primary dimensions to validate its deployment readiness:

End-to-End Latency: Measured continuously from the moment of far-field speech acquisition to the final state update in the risk-governance FSM. The hard real-time constraint for main control room applications was defined as a maximum delay of 200 ms.

Concurrency Capacity: Defined as the maximum number of parallel audio streams the edge node can process simultaneously without violating the 200 ms latency boundary or dropping acoustic frames.

Alert Reliability (False-Positive Rate): Evaluated to determine the framework's resistance to environmental noise and non-target speech. The proposed dual-gating architecture is benchmarked against a conventional baseline pipeline (which omits the ECAPA-TDNN speaker gate) using a standardized False-Positive Rate (FPR) statistic to quantify the reduction in nuisance alerts.

Table 2: Performance and reliability evaluation of the edge-native framework.

Metric	Proposed Framework (Edge)	Baseline (Cloud-based / No Gating)	Interpretation
End-to-end latency	185 ms	420 ms	Measured from audio capture to risk FSM state update. Meets real-time response constraints (<200ms).
Max concurrency	12 concurrent streams	N/A	Sustained parallel channels on a single NVIDIA RTX 4090 edge node without frame dropping.
False-positive index	11.2%	38.5%	Evaluated under 65dB ambient machinery noise. The dual-gating front end significantly reduces false alerts.

EXPERIMENTAL RESULTS AND DISCUSSION

Edge Processing Performance: Latency and Concurrency

The primary operational constraint in main control rooms is timely anomaly detection without relying on cloud infrastructure. Our evaluation demonstrates that the proposed edge-native architecture achieves an average end-to-end latency of 185 ms. This measurement encompasses the full pipeline: FRCRN enhancement, ECAPA-TDNN speaker gating, parallel analytics (affective, semantic, and fatigue), and the final risk fusion FSM. Compared to a conventional cloud-offloaded baseline (420 ms latency, heavily dependent on network jitter), the local deployment ensures deterministic real-time response. Furthermore, the single workstation-class edge node successfully

sustained 12 concurrent high-fidelity audio streams without violating the 200 ms latency boundary, proving its scalability for multi-operator control rooms.

Alert Reliability and Dual-Gating Efficacy

In high-stakes industrial environments, high false-positive rates lead to alarm fatigue. We evaluated the alert reliability under simulated control room acoustics characterized by 65 dB background machinery noise and intermittent speech overlap. The proposed system achieved a False-Positive Rate (FPR) of 11.2%, a substantial improvement over the 38.5% FPR observed in the baseline pipeline lacking the dual-gating front end. By filtering out unauthenticated noise and cross-speaker leakage at the ECAPA-TDNN gate before invoking the parallel cognitive-state analytics, the framework isolates the target operator's signal and reduces spurious state transitions in the final risk model.

CONCLUSION

In this paper, we proposed an edge-native, privacy-preserving audio analytics framework specifically designed for contactless operator resilience monitoring in safety-critical main control rooms. By integrating a dual-gating acoustic front end (FRCRN enhancement and ECAPA-TDNN speaker gating) with parallel cognitive-state analytics, the system successfully extracts affective, semantic, and fatigue-oriented indicators without relying on intrusive wearable sensors.

Experimental evaluations demonstrate the operational viability of our proposed architecture. Deployed on a local edge node, the integrated pipeline achieves an average end-to-end latency of 185 ms and sustains up to 12 concurrent streams, strictly satisfying the real-time constraints of industrial environments. Furthermore, in highly noisy conditions (65dB background noise), the dual-gating mechanism mathematically isolates the target speaker, significantly reducing the false-positive alert rate to 11.2% compared to conventional non-gated baselines.

Ultimately, this framework provides a deployable, low-burden solution that preserves data sovereignty while effectively mitigating supervisory alarm fatigue. Future work will focus on longitudinal deployment in operational control rooms to calibrate the short-term acoustic fatigue proxy against clinical cognitive baselines and long-term operator performance metrics.

REFERENCES

Desplanques, B., Thienpondt, J. and Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In Proceedings of Interspeech 2020, pp. 3830–3834.

Gao, Z., Zhang, S., McLoughlin, I. and Yan, Z. (2022). Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition. In Proceedings of Interspeech 2022, pp. 2063–2067.

Gomez-Zaragoza, L., del Amor, R., Castro-Bleda, M.J., Naranjo, V., Alcaniz Raya, M. and Marin-Morales, J. (2024). EMOVOME: A Dataset for Emotion Recognition in Spontaneous Real-Life Speech. arXiv preprint arXiv:2403.02167.

Hrnciar, M., Evin, E., Bures, M. and Bednarik, J. (2024). Prediction of Mental Fatigue in Main Control Room Operators Based on Facial Images. International Journal of Environmental Research and Public Health, 21(8), 1034.

Katav, A., Moshe, Y. and Cohen, I. (2025). A Framework for Robust Speaker Verification in Highly Noisy Environments Leveraging Both Noisy and Enhanced Audio. arXiv preprint arXiv:2508.18913.

Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S. and Chen, X. (2024). emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. Findings of the Association for Computational Linguistics: ACL 2024, pp. 15653–15669.

Mahalingam, S., Boominathan, P. and Arunachalam, R. (2021). Cepstral measures to analyze vocal fatigue in individuals with hyperfunctional voice disorder. Journal of Voice, 35(2), 230–239.

McEwen, B.S. and Stellar, E. (1993). Stress and the individual: Mechanisms leading to disease. Archives of Internal Medicine, 153(18), 2093–2101.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology, 14(4), 261–292.

Nanjundeswaran, C., Jacobson, B.H., Gartner-Schmidt, J. and Verdolini Abbott, K. (2015). Vocal Fatigue Index (VFI): Development and Validation. Journal of Voice, 29(4), 433–440.

Zhao, S., Ma, B., Watcharasupat, K.N. and Gan, W.-S. (2022). FRCRN: Boosting Feature Representation Using Frequency Recurrence for Monaural Speech Enhancement. arXiv preprint arXiv:2206.07293.